Video Action Recognition for Lane-Change Classification and Prediction of Surrounding Vehicles

Mahdi Biparva, David Fernández-Llorca, *Senior Member, IEEE*, Rubén Izquierdo Gonzalo, and John K. Tsotsos, *Fellow, IEEE*

***Abstract*—In highway scenarios, an alert human driver will typically anticipate early cut-in/cut-out maneuvers of surrounding vehicles using visual cues mainly. Autonomous vehicles must anticipate these situations at an early stage too, to increase their safety and efficiency. In this work, lane-change recognition and prediction tasks are posed as video action recognition problems. Up to four different two-stream-based approaches, that have been successfully applied to address human action recognition, are adapted here by stacking visual cues from forward-looking video cameras to recognize and anticipate lane-changes of target vehicles. We study the influence of context and observation horizons on performance, and different prediction horizons are analyzed. The different models are trained and evaluated using the PREVENTION dataset. The obtained results clearly demonstrate the potential of these methodologies to serve as robust predictors of future lane-changes of surrounding vehicles proving an accuracy higher than 90% in time horizons of between 1-2 seconds.**

***Index Terms*—Video action recognition, lane change prediction, surrounding vehicles, autonomous vehicles.**

I. Introduction

ONE of the closest and most plausible scenarios in the adoption of the autonomous vehicles is autonomous navigation at SAE L3 (chauffeur) or L4 (autopilot) on highways, both for passenger and freight transport. This is mainly due to the maturity of one of the first driver assistance technologies: Adaptive Cruise Control (ACC) systems. They were introduced in the early 1990 s and are present in a wide range of passenger vehicles today [1]. ACC systems focus on maintaining a desired speed selected by the driver or maintaining the distance between the car in front and the ego car. Newer versions introduced Stop & Go functionality. But the steering wheel must be controlled manually (L1).

The next step in automation were the Traffic Jam Assist (TJA) and the Traffic Jam Chauffeur (TJC) that combines the ACC Stop & Go and Lane Keeping Assist functions to control the steering wheel, speed, acceleration and braking of the vehicle in traffic jams up to speeds typically below 60 km/h. TJA is usually considered as L2 and TJC as L3 [2].

Finally, the most advanced automation systems to date are the Highway Chauffeur (HC) and the Highway Autopilot (HA), which includes the management of complex maneuvers such as deciding to change lanes to overtake, enter a slower lane or even exit the highway. HC is mostly considered as L3 and HA as L4 [2].

In all previous systems, from the simple ACC to the most sophisticated HA, the most critical, and challenging, highway scenarios are the cut-in and cut-out ones, specially for high speeds. In the cut-in scenario, a car from one of the adjacent lanes merges into the lane just in front of the ego car. In the cut-out scenario, a car in front leaves the lane abruptly to avoid a slower vehicle, or even stopped, ahead. Since 2018, the performance of these assistance or chauffeur commercial systems operating under these two critical traffic scenarios is being tested by Euro NCAP [3]. Although there are abnormal behaviors that can also lead to critical situations on highways, and that are of interest to driving automation systems, such as sudden stops, abnormal trajectories or collisions, the amount of data available is still very limited to develop, validate and certify potential approaches.

An alert driver will typically anticipate cut-in and cut-out maneuvers, even over long distances, using only visual cues, reduce speed accordingly, or even change lanes through the use of the steering wheel. An automated system must also be able to anticipate these situations at an early stage. To do so, it is necessary to endow new automated systems with the ability of predicting the motions of surrounding vehicles, such as lane-keeping and lane-change, and thus improving driving performance significantly in terms of safety, comfort, and even environmental sustainability [4], [5].

The first Lane Departure Warning (L0) or Lane Change/Lane Keeping Assist systems (L1) were designed to detect, or

Manuscript received 12 November 2021; revised 23 February 2022; accepted 30 March 2022. Date of publication 4 April 2022; date of current version 24 October 2022. This work was supported in part by the Spanish Ministry of Science, Innovation and Universities under Salvador de Madariaga Mobility Grants PRX18/00155, DPI2017-90035-R, and PID2020-114924RB-I00, in part by the Community Region of Madrid under Grant 2018/EMT-4362 SEGVAUTO 4.0-CM, in part by the Air Force Office of Scientific Research USA under Grant FA9550-18-1-0054, in part by Canada Research Chairs Program under Grant 950-231659, and in part by Natural Sciences and Engineering Research Council of Canada under Grant RGPIN-2016-05352. *(Corresponding author: David Fernández-Llorca.)*

Mahdi Biparva and John K. Tsotsos are with the Department of Electrical Engineering and Computer Science, York University, Toronto, ON M3J 1P3, Canada (e-mail: mhdbprv@cse.yorku.ca; tsotsos@eecs.yorku.ca).

David Fernández-Llorca is with Computer Engineering Department, University of Alcalá, Alcalá de Henares, 28805 Madrid, Spain, and also with European Commission - Joint Research Center, 41092 Seville, Spain (e-mail: david.fernandezl@uah.es).

Rubén Izquierdo Gonzalo is with Computer Engineering Department, University of Alcalá, Alcalá de Henares, 28805 Madrid, Spain (e-mail: ruben.izquierdo@uah.es).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TIV.2022.3164507.

Digital Object Identifier 10.1109/TIV.2022.3164507

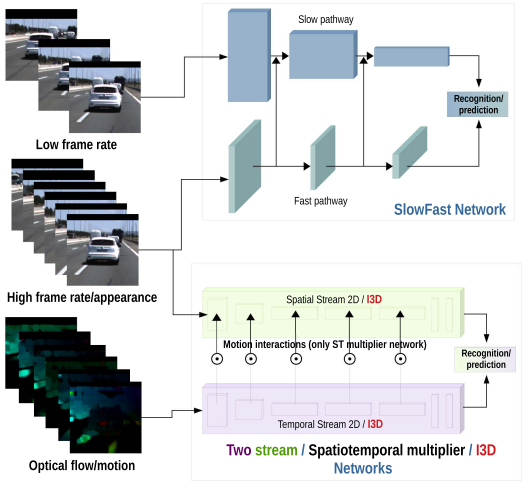


Fig. 1. Overview of the proposed video action recognition approaches for lane change recognition and prediction of surrounding vehicles, including Two-Stream Network, Two-Stream Inflated 3D ConvNet, Spatiotemporal Multiplier Network and SlowFast Network.

even predict, lane departure of the ego vehicle by combining visual cues (lane markings and lane texture) and vehicle-state information (CAN bus) [6]. Although there is some ambiguity in the available literature, ego-vehicle lane change detection systems differ considerably from surrounding vehicle lane change detection systems. The requirements for sensors are very different, as are the applicable methodologies. For example, the pixel resolution available to detect the position of vehicles relative to their lane is much lower. Solutions require vehicle detection and tracking. Relative distance and speed measurements require radar or LiDAR type range sensors, and yet, uncertainty of the measurements is much more relevant. Vehicle-state information (e.g., accelerations), can only be accurately obtained via V2V communications [7].

To deal with lane-change prediction of surrounding vehicles, in this paper we pose the problem as an action recognition problem using visual information from cameras. The idea behind our proposal is to use the same source of information (visual cues) and the same type of approach (action recognition) that drivers use to anticipate these maneuvers. By using a spatio-temporal model based on image sequences (i.e., continuous visual cues) our approach implicitly includes positional, contextual, and symbolic information, such as turn or brake indicators. Although there are some drivers who do not use them (i.e., a system based solely on their detection would not be effective), in general they are a very valuable source of visual information.

Significant progress has been made in video-based human action recognition and prediction during the last years [8]. Action recognition and prediction involves managing spatial and temporal information (sequence of images). Among the different methodologies, we focus our efforts in the following two-stream-based approaches (see Fig. 1):

- *Two-Stream Convolutional Networks [9]:* a classical architecture that contains a spatial network and a temporal network (two streams), which are used for modeling static information in still frames and motion information in optical flow images, respectively.
- *Two-Stream Inflated 3D Convolutional Networks (I3D) [10]:* an extension of the classical two-stream architecture which expand filters and pooling kernels into 3D, leading to very deep, naturally spatiotemporal classifiers.
- *Spatiotemporal Multiplier Networks [11]:* a two-stream architecture that combines appearance and motion pathways and allows interaction between them by injecting cross-stream residual connections.
- *SlowFast Networks [12]:* a two-stream architecture involving a slow pathway that operates at low frame rate to capture spatial semantics and a fast pathway that operates at high frame rate to capture motion at fine temporal resolution.

Although there are other works focused on learning spatiotemporal features for video activity recognition, the selected approaches are a good example of the evolution of the two-stream-based systems, including the first successful proposal [9] and the one ranked first [12] in the AVA Challenge 2019 [13]. Beyond our previous preliminary work [14], to the best of our knowledge, this is the first proposal using video action recognition approaches to deal with lane-change recognition and prediction of surrounding vehicles for automated vehicles.

To validate these approaches in this context, we make use of The PREVENTION dataset [15] which provides a large number of accurate and detailed annotations of vehicles categories, trajectories and events (including left/right lane changes, among others). More than 356 minutes, 4 M vehicle detections and 3 K trajectories are available, with data collected from LIDAR, radar and camera sensors, from surrounding vehicles up to a range of 80 meters. Contours and bounding boxes are available as raw output detections, as well as a temporary integration of the detections.

The aforementioned architectures are adapted to deal with lane change action recognition and prediction. An extensive evaluation is performed in this paper. The amount of context information needed to model the interactions between different vehicles and other features implicitly included in the appearance, such as the number of lanes or the road curvature, is studied using different sizes for the regions of interest. The ability of the networks to perform action recognition and prediction is assessed using different time horizons and training strategies. The obtained results clearly validate the use of these type of approaches to solve the lane-change prediction problem of surrounding vehicles.

The remainder of the paper is organized as follows. In Section II, the related work is presented, whereas Section III is an overview of the problem formulation. In Section IV the implementation of the action recognition approaches is described. In Section V the evaluation metrics, and the performance of the different approaches are assessed. The final conclusions and future work are given in Section VI.

II. Related Work

Most of the available work on lane-change recognition and prediction focuses on in-vehicle detection. However, as stated before, the nature of the problem is considerably different,

so we limit our analysis of lane-change detection of other vehicles, and more specifically, within the context of the highway scenario. Vehicle and lane markings detection and tracking [16] are necessary conditions. However, it is reasonable to consider them as separated problems that are independent of the maneuver recognition system.

Three levels of analysis will be considered. First, we will review the type of input features used. Second, we will focus on the different types of methodologies. Finally, we will describe the available datasets and their main features.

A. Input Variables

Most of the works analyzed are based on the use of physical variables that define the relative dynamics of the vehicle with other vehicles and with its environment [17]–[31]. Some of these variables are lateral and longitudinal position (distances), velocity, acceleration, timegap, heading angle and yaw rate. These variables are usually obtained and processed in a multi-modal fashion, by fusing data from onboard sensors such as cameras and range sensors (radar and/or LiDAR). Errors and uncertainties in the estimation of these variables from the raw data lead to additional limitations. We can expect reasonably accuracy when measuring the position and relative velocities of other vehicles using onboard sensors. However, it is unrealistic to handle accurate measurements of variables such as lateral and longitudinal accelerations, yaw angle, or yaw rate. As an example, we refer to [32] to see the intrinsic difficulty of obtaining accurate speed measurements from static cameras. Sensor uncertainties are intrinsically modeled in some approaches [17], [30], but even so, we cannot expect them not to affect predictions. In some cases it is assumed that these variables will be available via V2V communications [20], but this scenario requires a 100% penetration rate, and in that case, predicting the intentions of other vehicles would be unnecessary as the vehicles could transmit their intentions. In addition, V2V communications pose a number of additional problems to consider [33]. In any case, we are still far from this scenario.

Context cues are also introduced, including road-level features such as the curvature and speedlimit [19], [21], [30], distance to the next highway junction [23], number of lanes [29], etc., as well as lane-level features such as type of lane marking or the distance to lane end [23]. These variables are inferred and processed from camera sensors, localization systems and enhanced digital maps, and are also subject to errors and uncertainties that will affect detection and prediction performance.

The number of proposals making use of appearance features to perform lane change recognition or prediction is surprisingly low (excluding vehicle and lane markings detection which are common features in all approaches), especially considering that human drivers do not use the physical variables mentioned above to anticipate lane changes from other vehicles but visual cues. In [34] the position of the vehicle bounding box in the image (in pixels) is used, but no appearance features are extracted. This approach is very sensitive to camera position, orientation and settings. In [30], two variables manually selected from the appearance, i.e., state of turn indicators and state of brake indicators, are used. Likewise, these variables are obtained from a specific detection system that involves errors and uncertainties that will affect later stages. But the main limitation of systems that explicitly seek to detect turn indicators is that in many cases drivers do not make use of them when changing lanes. In our previous works [35], [36] regions of interest (ROIs) are generated for each vehicle detection, including local information around the vehicle, and appearance features are extracted using a GoogLeNet pre-trained on ImageNet. Using the raw image data (appearance) as input to the lane change detection and prediction system is challenging, but has the benefit of not requiring intermediate detection steps that can introduce additional errors and uncertainties.

B. Methodologies

As suggested by [37] vehicle motion modeling and prediction approaches can be classified into three different levels: physical-based, where predictions only depend on the laws of physics, maneuver-based, where the future motion of a vehicle depends on the driver maneuver, and intention-aware, where predictions take into consideration inter-dependencies between vehicles. Note that, as a chicken-egg problem, on the one hand, lane-change recognition can be addressed using the trajectory estimated by any of the motion models [38], and on the other hand, the prediction of the trajectories of surrounding vehicles can be estimated more accurately if the lane-change intention recognition is available.

Some proposals are intention-aware in their nature. For example, by using graphical models such as Bayesian Networks [17], [23], [30] or Structural Recurrent Neural Networks [29], or by using convolutional social pooling in an LSTM encoder-decoder architecture [28]. However, in most cases, inter-dependencies between vehicles are modeled by extracting relative physical features (distances, velocities or time-gaps) [19], [21], [25], [27] or by generating compact representations that encode the relative positions of all vehicles on the scene [26], [35], [36]. A considerable number of previous works do not take into consideration the interaction between vehicles [18], [20], [22], [31], [34].

Many approaches to lane-change recognition and prediction address the problem using generative-based solutions, including Naïve Bayes Classifiers [19], Bayesian Networks [17], [23], [30], and Hidden Markov Models [20]. Others make use of discriminative solutions such as case-based reasoning [18], Random Decision Forest [21], traditional Neural Networks [22], [24], Support Vector Machines [24], [25], [34], Gaussian Process Neural Networks [31], and feedforward Convolutional Neural Networks [26], [35], [36]. Finally, some other approaches are based on the use of Recurrent Networks including vanilla LSTM [35] and LSTM encoder-decoder [27] and multi-modal [28] architectures. Consequently, two-stream architectures have not been proposed so far to perform lane change detection and prediction.

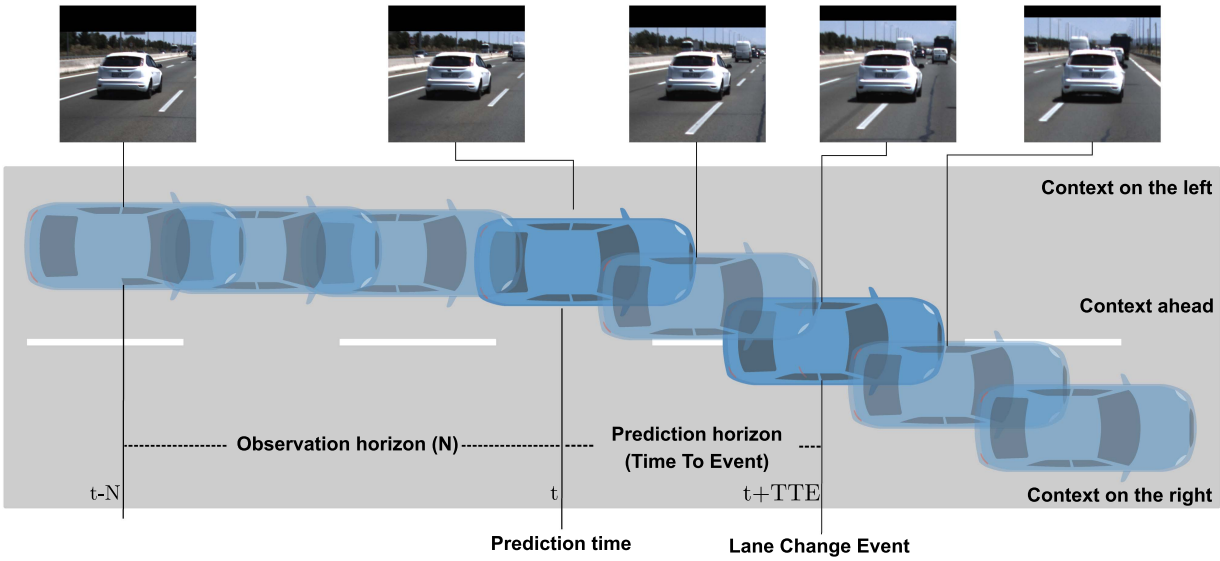


Fig. 2. Problem formulation: observation horizon (N), and time to event (TTE). The lane change event is labeled as the frame where the middle of the rear bumper is located just over the lane markings. This is the criterion established in PREVENTION dataset [15].

C. Datasets

In order to train learning-based approaches and validate the quality of the proposed solutions, available datasets play a fundamental role. Two type of recording setups are usually proposed depending on the location of the sensors. First, we have datasets captured from the infrastructure using cameras installed on buildings, such as NGSIM HW101 [39] or NGSIM I-80 [40] datasets, or cameras on-board drones, such as HighD [41], inD [42] or INTERACTION [43] datasets. Although these datasets are very valuable for understanding and assessing the motion and behavior of vehicles and drivers under different traffic scenarios, they are not fully applicable for on-board detection applications.

Second, other datasets provide road data with sensors on-board vehicles. In this line, the PKU dataset [44] was released in 2017 by Peking University and the PSA Group, containing 170 minutes of data gathered using a vehicle equipped with 4 2D-LiDARs covering a region of 40 meters around the vehicle. It does not contain information regarding the road lane markings, the number of road lanes, or the relative positioning of the ego-vehicle. In 2018, the ApolloScape dataset [45] was released by Baidu Research, containing data obtained in urban environments from 4 cameras and 2 Laser scanners using a vehicle driving at 30 km/h. It is currently one of the most complete datasets in the state-of-the-art but it does not contain radar data, making detections more sensitive to failure in adverse weather conditions and highway scenarios. In addition, it does not provide labeled tracking information (IDs and tracklets) for all detected objects. In 2019, the PREVENTION dataset [15] was released containing data from 3 radars, 2 cameras and 1 LiDAR, covering a range of up to 80 meters around the ego-vehicle (up to 200 meters in the frontal area). Road lane markings are included and the final position of the vehicles is provided by fusing data from the three type of sensors.

III. Problem Formulation

We define lane change prediction as a multi-classification problem in which the goal is to determine whether a vehicle i will make a left or right lane-change (LLC, RLC) or remain in its lane (no lane change) given the observed context up to some time N. As can be seen in Fig. 2, the lane-change event is defined as the time when the center of the rear bumper is just above the lane markings. Therefore, cases with small lateral displacements, lateral oscillations, or aborted lane change maneuvers (including unsafe or aggressive behaviors) are contained in the no lane change class (NLC). Although these are difficult cases that often result in false positives, false lane change detection and prediction would not be as critical for the context of predictive driving automation systems as they can anticipate dangerous situations in which the safest control actions would be the same as in the case of actual lane changes.

The observation horizon or time window will contain a set of N images that will be stacked according to the activity recognition method used.

Then, the problem can be posed as a *classification* or *prediction* problem based on the value of the Time to Event (TTE), or prediction horizon, as follows:

- Lane-change *classification:* when $TTE = 0$. That is, the observation horizon contains part of the lane change maneuver itself for the LLC and RLC classes.
- Lane-change *prediction:* when $TTE > 0$. Depending on the TTE value, the observation horizon will contain more or less information of the actual lane change maneuver for LLC and RLC classes. For very high TTE values the maneuver may not even have started. Still, contextual or symbolic information can help anticipate lane changes in these cases.

We will examine the effects of TTE or prediction horizon and observation duration (N) on the accuracy of lane-change classification and prediction.

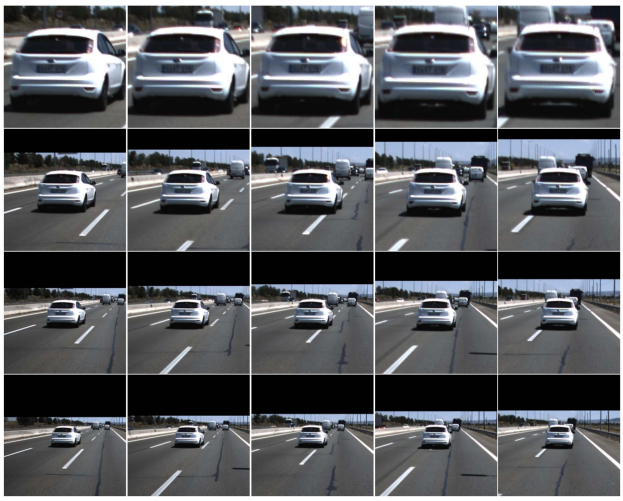


Fig. 3. ROI sizes. From upper row to lower row: x1, x2, x3 and x4. The vehicle is always centered. Zero-padding is applied when needed.

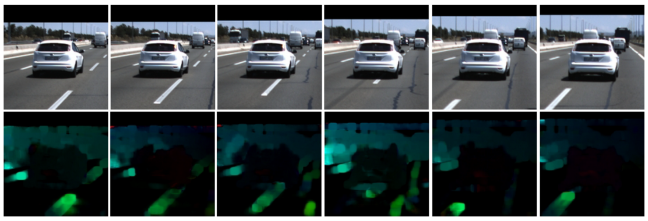


Fig. 4. Example of dense optical flow computation.

The prediction relies on visual cues that are computed from regions of interest (ROI) extracted from the contour labels provided in the PREVENTION dataset. Four different ROI sizes are considered: $\times 1$, $\times 2$, $\times 3$ and $\times 4$ the size of the square bounding box around the vehicle contour (see Fig. 3). Zero-padding is used when the ROI exceeds the limits of the image. The size of the ROI modulates the amount of context information being considered in the input data stream. Thus, $\times 1$ mostly contains information related with the vehicle appearance, while $\times 4$ incorporates a large amount of front and side context information. For ROI sizes of $\times 3$ and $\times 4$ the approach can be considered interaction-aware since the image contains information regarding cars in the same or adjacent lanes. Other variables relevant for lane change prediction such as the number of lanes, or road curvature, are implicitly included in the context information.

Since the vehicle is always centered in the ROI, dense optical flow (from the motion stream) should be interpreted as a way of measuring the movement of the context (infrastructure and other vehicles) around the detected vehicle. As shown in Fig. 4, the optical flow is low in the region where the vehicle is, while it is more predominant around it.

IV. Video Activity Recognition & Prediction

The sequence of stacked images or regions of interests, can naturally be decomposed into spatial and temporal components. The spatial part, in the form of individual region appearance, carries information about the vehicle itself (e.g., light indicators or brake lights) and the context around it (road, lane markings and surrounding vehicles). The temporal part, in the form of motion across frames, conveys the movement of the observer (onboard camera) w.r.t. to the road, and the surrounding vehicles. In order to handle a canonical view for the motion stream, all the regions are generated around the contour of the vehicle so the vehicle is always centered in the region of interest (the size will vary depending on the relative distance w.r.t. the ego vehicle). We consider four video activity recognition approaches: Disjoint Two-Stream Convolutional Networks (TS) [9], Two-Stream Inflated 3D Convolutional Networks (I3D) [10], Spatiotemporal Multiplier Networks (STM) [11] and SlowFast Networks (SF) [12].

A. Disjoint Two-Stream Convolutional Networks

A two-stream ConvNet architecture which incorporates and fuses spatial and temporal information is defined. The structure of the ConvNets for both streams is the same, including 5 convolutional layers and 3 fully connected layers, with the parameters depicted in Fig. 5. The last fully connected layer is defined with 3 outputs regarding the three classes defined: left lane change (LLC), right lane change (RLC), and no lane change (NLC).

The dense optical flow is computed using polynomial expansion [46]. The spatial stream ConvNet is pre-trained using ImageNet and the temporal ConvNet using multi-task learning using UCF-101 and HMDB-51. All hidden layers use the rectification (ReLU) activation function. Max-pooling is performed over 3×3 spatial windows with stride 2.

B. Two-Stream Inflated 3D Convolutional Networks

The natural approach to deal with video modeling is to use 3D convolutional neural networks. These are like standard convolutional networks, but with spatio-temporal filters that generate a hierarchical representation of spatio-temporal data. These are more complex architectures with a higher number of parameters that cannot easily benefit of pre-training strategies. We adopt the approach presented in [10] which starts with a 2D architecture and inflates all the filters and pooling kernels endowing them with an additional temporal dimension. Each 3D network is implemented with 8 convolutional layers, 5 pooling layers and 2 fully connected layers at the top. Batch normalization is applied after all convolutional and fully connected layers. The 3D filters are bootstrapped from pre-trained ImageNet models by repeatedly copying an image into a video sequence. A two-stream configuration is used (see Fig. 6), learning temporal patterns from the appearance stream, but enhancing its performance by including the motion stream. The inputs to the model are short 16-frames sequences.

C. Spatiotemporal Multiplier Networks

The original two-stream architecture only allows the two processing streams (spatial and motion) to interact via late fusion of their respective softmax predictions. This way, the architecture does not support the learning of truly spatiotemporal features, since the loss of both streams is backpropagated independently

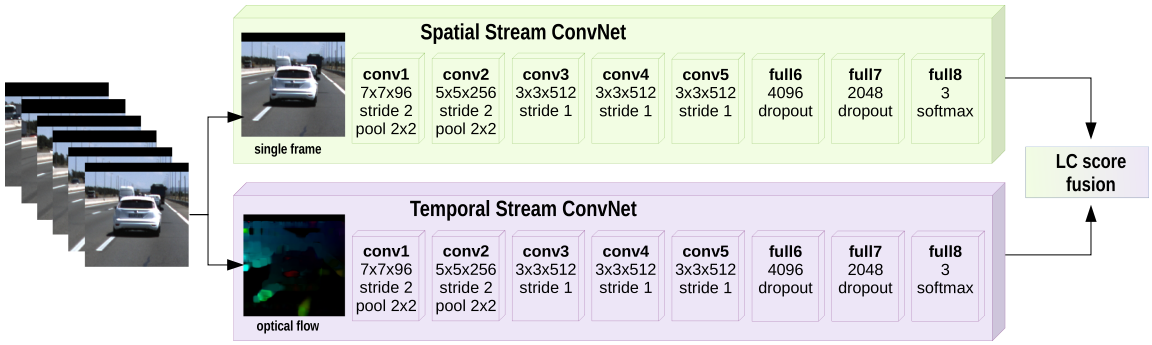


Fig. 5. Disjoint two-stream architecture for lane change classification and prediction.

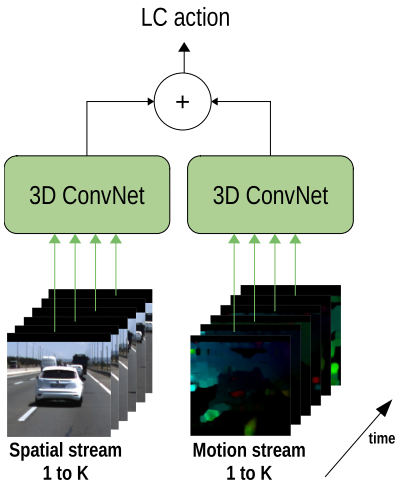


Fig. 6. Two-stream inflated 3D ConvNet for lane change classification and prediction.

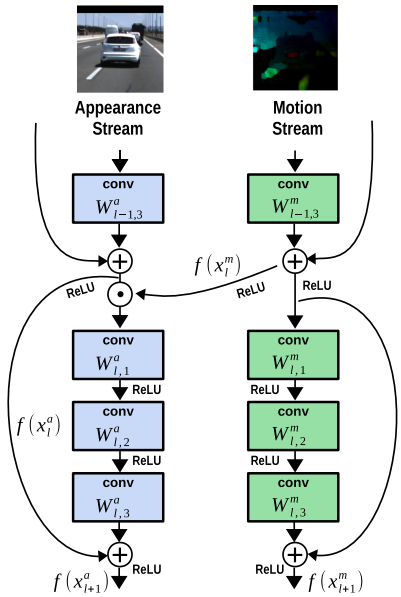


Fig. 7. Multiplicative residual gating from the motion stream to the appearance stream.

without any type of interaction. Learning spatiotemporal features requires the appearance and motion paths to interact earlier on during the forward pass. This interaction can be relevant for the classification and prediction of lane change maneuvers that have similar appearance or motion patterns and can only be inferred by the combination of two (e.g., vehicles that do not change lanes but have their turn indicators on). To address this limitation, it is possible to inject cross-stream residual connections using Residual Networks (ResNets) [47] as the general architecture for the spatial and the temporal streams.

In [11], different cross-stream connections were studied, including two types of connections (direct or into residual units), two fusion functions (additive or multiplicative), and different streams directions (unidirectional from the motion into the appearance, conversely and bidirectional), being the multiplicative residual connection from the motion path into the appearance stream the one providing the superior performance.

As can be observed in Fig. 7, the multiplicative interaction can be formulated as:

$$\hat{x}^a_{l+1} = f(x^a_l) + \mathcal{F}\left(x^a_l \odot f(x^m_l), W^a_l\right) \tag{1}$$

where x^a_l and x^m_l are the inputs of the l-th layers of the appearance and motion paths respectively, while W^a_l represents the weights of the l-th layer residual unit in the appearance stream and $\odot$ corresponds to elementwise multiplication.

Better temporal support is also provided by injecting 1D temporal convolutions layers into the network [11]. ResNet50 model is used for both streams, including batch normalization and ReLU activation function after each convolutional block.

D. SlowFast Networks

One of the most successful video action recognition approaches is the so called SlowFast network [12]. It can be considered as a two-stream approach, although motion pathway is not directly used. Instead, one stream (slow) is designed to capture semantic information given by a few sparse images operating at low frame rates and slow refreshing speed, and a second stream (fast) is responsible for capturing rapidly changing motion by operating at high temporal resolution and fast refreshing speed. The two pathways are fused by lateral connections. The temporal stride used in the Slow pathway is $\tau = 16$ and the frame rate ratio between the Fast and Slow streams is $\alpha = 8$. The ratio of channels of the Slow stream with respect to the Fast one is

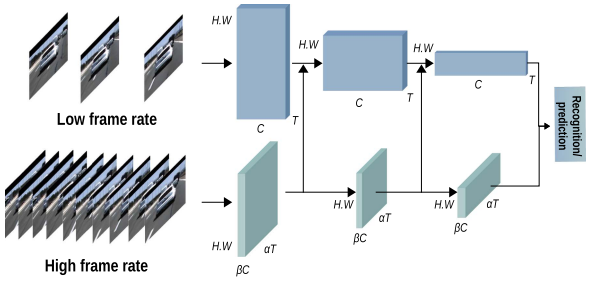


Fig. 8. SlowFast network for lane change recognition and prediction. The fast stream is lightweight by using a fraction $\beta = 1/8$ of channels.

TABLE I
MAIN STATS OF THE DATASET. NLC/LLC/RLC: NO/LEFT/RIGHT LANE-CHANGE

	NLC	LLC	RLC
# of sequences	3110	342	438
avg. # of frames	50.9	96.8	80.1

defined as $\beta = 1/8$ (see Fig. 8). The network is defined with one convolutional layer, five residual blocks and one fully connected layer adapted to the number of classes as in [12]. Since optical flow is not computed, the architecture can be learned end-to-end from the raw data.

V. EXPERIMENTS

A. Dataset Description

Table I summarizes the details of the dataset. The input size for both streams is 112×112. The 85% of the samples are used for training and the remaining 15% for validation.

B. Evaluation Parameters and Models

The following parameters have been evaluated during the experiments:

- ROI sizes: x1, x2, x3 and x4.
- Observation horizon: 20 frames (2 seconds), 30 frames (3 seconds) and 40 frames (4 seconds).
- Time-to-event (prediction horizon): 0 (no prediction), 10 (1 s) and 20 (2 seconds).

The evaluated video recognition models are the Disjoint Two-Stream ConvNet (**Disjoint**), the Two-Stream Inflated 3D ConvNet (**I3D**), the Spatiotemporal Multiplier ConvNet (**ST**), and the SlowFast ConvNet (**SF**). A basic model which implements the appearance stream of the Disjoint architecture (upper pathway in Fig. 5) is used as the baseline (**Baseline**). In all cases, the specific architecture of the models and the hyper-parameters used for training are those reported as optimal by the authors.

C. Metrics

As a multi-class problem (with 3 classes), we use the categorical entropy loss function for the training. For evaluating the results, we consider the accuracy as the main variable to assess the performance of the two evaluated methods and the corresponding parameters, i.e., the number of true positives for the three classes divided by the total number of samples (arithmetic mean of precision for all classes). In addition, we evaluate precision and recall for all classes in confusion matrices.

TABLE II
LANE-CHANGE CLASSIFICATION ($TTE = 0$) ACCURACY (%)

		ROI size			
Method	Obs. Horizon	x1	x2	x3	x4
Baseline	20	83.41	83.25	**85.35**	84.06
	30	81.96	83.25	82.61	85.19
	40	81.80	82.45	81.32	81.80
Disjoint	20	83.22	86.18	86.26	87.43
	30	83.55	86.69	86.84	86.68
	40	84.97	87.69	**89.46**	88.79
I3D	20	82.45	**86.47**	85.99	85.67
	30	82.13	83.74	83.90	84.06
	40	82.13	83.09	81.80	82.29
ST	20	83.39	85.03	86.51	86.16
	30	84.38	84.70	85.36	84.73
	40	86.02	87.83	**90.30**	89.64
SF	20	88.89	89.69	**90.98**	89.37
	30	88.57	89.53	88.24	89.69
	40	86.96	89.05	89.53	90.34

D. Lane-Change Classification Results

In Table II we depict the lane-change classification (i.e., with $TTE = 0$) accuracy of all action recognition approaches over the validation set.

Regarding the ROI sizes we can state the following conclusions. By using just the ROI fitted to the bounding box, the results are surprisingly reasonable, considering that almost no context and interaction are available. In general, the higher the ROI size, the better the accuracy (with the exception of the I3D model), although adding more context from $x3$ to $x4$ decreases the performance for most cases and observation horizons. This can be explained by the fact that the observation horizons already incorporate context into the spatial, motion and slow streams, so using a higher ROI is not reflected in a better performance.

The effect of the observation horizon depends on the model. For example, the Disjoint and the Spatiotemporal models yield the best classification performance with the longest observation horizon (4 seconds). However, the I3D and the SlowFast architectures have a higher accuracy with the shortest observation horizon (2 seconds). For the classification task, the last frames are the most informative, and these models (I3D and SlowFast) seem to take better advantage of this information without the need for a larger observation horizon.

The best classification results, 90.98%, are provided by the SlowFast model with a ROI size of $x3$ and an observation horizon of 2 seconds, followed by the Spatiotemporal Multiplier Network, 90.30% with a ROI size of $x3$ and an observation horizon of 4 seconds.

E. Lane-Change Prediction Results

The ability of all methodologies to predict the future lane-change maneuverer of target vehicles is evaluated using an

TABLE III
LANE-CHANGE PREDICTION ACCURACY (%). OBSERVATION HORIZON = 20 FRAMES (2 SECONDS)

		ROI size			
Method	TTE	x1	x2	x3	x4
Baseline	10	82.63	82.95	83.44	82.79
	20	82.00	81.67	82.79	**83.61**
Disjoint	10	84.05	84.54	85.20	85.36
	20	85.20	88.82	**91.02**	90.92
I3D	10	81.33	83.28	83.60	83.60
	20	81.01	81.67	**83.93**	83.61
ST	10	84.70	85.69	85.20	86.51
	20	86.84	90.30	91.45	**91.94**
SF	10	85.23	**88.96**	88.64	87.99
	20*	85.27	83.31	83.61	83.61

*Inconclusive results due to GPU memory limitations.

TABLE IV
SPATIOTEMPORAL MULTIPLIER NETWORK CONFUSION MATRIX, OH=20, TTE=20, X4

	Target class			
Output class	**NLC**	**LLC**	**RLC**	**Precision**
NLC	476	5	6	97.7%
LLC	8	33	11	63.5%
RLC	11	8	50	72.5%
Recall	96.2%	71.7%	74.6%	91.9%

observation horizon of 20 frames (2 seconds) and prediction horizons of 10 and 20 frames (1 and 2 seconds respectively). The results for all approaches are depicted in Table III.

Starting from the baseline, and with the exception of the I3D model, it is remarkable to see that predictions are better for longer prediction horizons, i.e., the obtained accuracy for $TTE = 20$ frames is generally higher than for a $TTE = 10$ images. This can be explained, in part, by the complexity of the models that better generalize with a more complex objective to learn. This effect is particularly visible with the Disjoint and ST models, where the accuracy is approximately 5% higher when predicting 2 seconds ahead than 1 s ahead.

Concerning the ROI size we can state that the larger the ROI the better the prediction accuracy, with no saturation effect from $x3$ to $x4$. The best performance when predicting lane-changes 1 s before they occur is obtained with the SlowFast model with a ROI size of $x2$, yielding an accuracy of 88.96%. For the larger prediction horizon, 2 seconds, the best model is the Spatiotemporal Multiplier network, which provides an accuracy of 91.94% with a ROI size of $x4$. Note that, the results of the SlowFast model for this case are inconclusive due to GPU memory problems. Whereas the mini-batch size for the other models was 32, the maximum size allowed with the SlowFast model was only 8. It is very likely that without this limitation, the SlowFast model would have provided even better results.

If we analyze the results further, we find that the predictions are closely linked to the number of samples available for each class. In fact, as shown in Table I, we have an unbalanced dataset which clearly affects the results. In Table IV we depict the confusion matrix for the best model (Spatiotemporal Multiplier Network) and the best parameters ($x4$, $TTE = 20$) including precision and recall.

As can be observed, the highest precision/recall ratio is obtained for the NLC class which represents almost the 80% of the samples. This correlation between the accuracy and the number of samples is also observed between the LLC and RLC classes, with a better precision/recall ratio for RLC class which contains 28% more samples than LLC. Some of the false positives for the LLC and RLC classes are due to instances of small lateral displacements, or aborted lane change maneuvers. However, the number of samples for these cases is not significant enough to draw further conclusions.

In any case, these are ones of the first prediction results so far using the PREVENTION dataset and the ability of the proposed two-stream multiplier network to predict lane-changes 2 seconds of anticipation is remarkable compared to the ability of humans trying to perform the same task (see [48] and [36] for more details on human performance).

VI. CONCLUSION AND FUTURE WORK

In this work, four video action recognition approaches have been adapted, trained and evaluated to deal with lane-change classification and prediction of target vehicles in highway scenarios using the labeled images and sequences available in the PREVENTION dataset. The anticipation of lane-changes is devised as an action recognition problem using visual cues from front view cameras, which is the same approach used by human drivers to predict these maneuvers. The Disjoint Two-Stream ConvNets (Disjoint), the Two-Stream Inflated 3D ConvNets (I3D), and the Spatiotemporal Multiplier ConvNets (ST) are based on two different pathways obtained from the same sequence of images: a spatial stream in the form of individual region appearance, and a motion stream in the form of dense optical flow across frames. The SlowFast ConvNet (SF) is based on two different pathways obtained from the appearance, but taken with two different sampling frequencies (one fast and one slow).

The influence of the context has been evaluated by using different ROI sizes, being the larger regions ($x3$ and $x4$ the original size of the vehicle) the ones providing the better classification and prediction results. For lane-change recognition, different observation horizons have been tested. Whereas the Disjoint and the ST models yield the best results with the longest observation horizon, the SF network provides the best recognition performance (91%) using visual cues from the shortest observation horizon evaluated, 2 seconds. The I3D model slightly outperforms the baseline, but its classification performance is lower than the rest.

The ability of most of these models (Disjoint, ST and SF) to predict lane-changes at $t + TTE$ is even better than their ability to classify them at t, which is a remarkable feature that can be partially explained by the high complexity of the models that provide better generalization with a more complex objective to learn. The best prediction results are obtained with the ST model with ROI size of $x4$ and observation horizon of 2 seconds, anticipating lane-changes 2 seconds earlier with an accuracy of 63.5% for left lane-changes, of 72.5% for right lane-changes and of 97.7% for no lane-change.

The presented video action recognition approaches are highly data-dependent, and the number of publicly available datasets is limited. As future works we plan to mitigate the problem with imbalanced classes of the dataset by applying resampling and data augmentation techniques, including generative adversarial networks (GANs) [49]. Other available and new datasets will be considered to alleviate this problem, to reduce potential bias and to analyze the generalization capabilities of action recognition methods. In cases where synchronized data from multiple sensors are available, the study of data fusion techniques will be addressed to overcome possible limitations of the vision-based method. GPUs memory limitations with models such as SlowFast will be also addressed. Other interesting topics to explore are the impact of lighting and weather conditions on system performance, and the potential ability of the system to explicitly detect and differentiate turn signals and brake lights (e.g., to assess whether or not a driver is correctly signaling an intention to change lanes). Finally, a real-time version will be devised to be tested in on-line real scenarios with our vehicle platforms.

REFERENCES

[1] L. Xiao and F. Gao, "A comprehensive review of the development of adaptive cruise control systems," *Veh. Syst. Dyn.*, vol. 48, no. 10, pp. 1167–1192, 2010.

[2] ERTRAC, "Connected automated driving roadmap," Mar. 8, 2019. [Online]. Available: https://www.ertrac.org/uploads/documentsearch/id57/ERTRAC-CAD-Roadmap-2019.pdf

[3] EuroNCAP, "2018 automated driving tests," Oct. 2018. [Online]. Available: https://www.euroncap.com/en/vehicle-safety/safety-campaigns/2018-automated-driving-tests/

[4] Y. Liu, Z. Wang, K. Han, Z. Shou, P. Tiwari, and J. Hansen, "Vision-cloud data fusion for ADAS: A lane change prediction case study," *IEEE Trans. Intell. Veh.*, to be published, doi: 10.1109/TIV.2021.3103695.

[5] D. Fernandez-Llorca and E. Gomez-Gutierrez, "Trustworthy autonomous vehicles," EUR 30942 EN, Publications Office Eur. Union, Luxembourg, no. JRC127051, 2021.

[6] J. McCall and M. M. Trivedi, "Video-based lane estimation and tracking for driver assistance: Survey, system, and evaluation," *IEEE Trans. Intell. Transp. Syst.*, vol. 487, no. 1, pp. 20–37, Mar. 2006.

[7] I. Parra, H. Corrales, N. Hernández, S. Vigre, D. F. Llorca, and M. A. Sotelo, "Performance analysis of vehicle-to-vehicle communications for critical tasks in autonomous driving," in *Proc. IEEE Intell. Transp. Syst. Conf.*, 2019, pp. 195–200.

[8] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *Int. J. Comput. Vis.*, vol. 130, pp. 1366–1401, 2022, doi: 10.1007/s11263-022-01594-9.

[9] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 568–576.

[10] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299–6308.

[11] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4768–4777.

[12] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6202–6211.

[13] C. Gu *et al.*, "AVA: A video dataset of spatio-temporally localized atomic visual actions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6047–6056. [Online]. Available: https://research.google.com/ava/challenge.html

[14] D. F. Llorca, M. Biparva, R. Izquierdo, and J. Tsotsos, "Two-stream networks for lane-change prediction of surrounding vehicles," in *Proc. IEEE Intell. Transp. Syst. Conf.*, 2020, pp. 1–6.

[15] R. Izquierdo, A. Quintanar, I. Parra, D. Fernández-Llorca, and M. A. Sotelo, "The prevention dataset: A novel benchmark for prediction of vehicles intentions," in *Proc. IEEE 22nd Int. Conf. Intell. Transp. Syst.*, 2019, pp. 3114–3121.

[16] S. Sivaraman and M. M. Trivedi, "Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 4, pp. 1773–1795, Dec. 2013.

[17] D. Kasper *et al.*, "Object-oriented Bayesian networks for detection of lane change maneuvers," *IEEE Intell. Transp. Syst. Mag.*, vol. 4, no. 3, pp. 19–31, Aug. 2012.

[18] R. Graf, H. Deusch, M. Fritzsche, and K. Dietmayer, "A learning concept for behavior prediction in traffic situations," in *Proc. IEEE Intell. Veh. Symp.*, 2013, pp. 672–677.

[19] J. Schlechtriemen, A. Wedel, J. Hillenbrand, G. Breuel, and K.-D. Kuhnert, "A lane change detection approach using feature ranking with maximized predictive power," in *Proc. IEEE Intell. Veh. Symp.*, 2014, pp. 108–114.

[20] P. Liu, A. Kurt, and U. Ozguner, "Trajectory prediction of a lane changing vehicle based on driver behavior estimation and classification," in *Proc. IEEE 17th Int. Conf. Intell. Transp. Syst.*, 2014, pp. 942–947.

[21] J. Schlechtriemen, F. Wirthmueller, A. Wedel, G. Breuel, and K.-D. Kuhnert, "When will it change the lane? A probabilistic regression approach for rarely occurring events," in *Proc. IEEE Intell. Veh. Symp.*, 2015, pp. 1373–1379.

[22] S. Yoon and D. Kum, "The multilayer perceptron approach to lateral motion prediction of surrounding vehicles for autonomous vehicles," in *Proc. IEEE Intell. Veh. Symp.*, 2016, pp. 1307–1312.

[23] M. Bahram, C. Hubmann, A. Lawitzky, M. Aeberhard, and D. Wollherr, "A combined model- and learning-based framework for interaction-aware maneuver prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 6, pp. 1538–1550, Jun. 2016.

[24] R. Izquierdo, I. Parra, J. M. Noz Bulnes, D. Fernández-Llorca, and M. A. Sotelo, "Vehicle trajectory and lane change prediction using ANN and SVM classifiers," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst.*, 2017, pp. 1–6.

[25] W. Yao *et al.*, "On-road vehicle trajectory collection and scene-based lane change analysis: Part II," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 1, pp. 206–2220, Jan. 2017.

[26] D. Lee, Y. P. Kwon, S. McMains, and J. K. Hedrick, "Convolution neural network-based lane change intention prediction of surrounding vehicles for ACC," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst.*, 2017, pp. 1–6.

[27] N. Deo and M. M. Trivedi, "Multi-modal trajectory prediction of surrounding vehicles with maneuver based LSTMs," in *Proc. IEEE Intell. Veh. Symp.*, 2018, pp. 1179–1184.

[28] N. Deo and M. M. Trivedi, "Convolutional social pooling for vehicle trajectory prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 1468–1476.

[29] S. Patel, B. Griffin, K. Kusano, and J. J. Corso, "Predicting future lane changes of other highway vehicles using RNN-based deep models," 2019, *arXiv:1801.04340v4*.

[30] J. Li, B. Dai, X. Li, X. Xu, and D. Liu, "A dynamic Bayesian network for vehicle maneuver prediction in highway driving scenarios: Framework and verification," *Electronics*, vol. 8, p. 40, 2019, doi: 10.3390/electronics8010040.

[31] M. Kruger, A. S. Novo, T. Nattermann, and T. Bertram, "Probabilistic lane change prediction using Gaussian process neural networks," in *Proc. IEEE 22th Int. Conf. Intell. Transp. Syst.*, 2019, pp. 3651–3656.

[32] D. F. Llorca *et al.*, "Two-camera based accurate vehicle speed measurement using average speed at a fixed point," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst.*, 2016, pp. 2533–2538.

[33] I. Parra, R. Izquierdo, J. Alonso, A. G. Morcillo, D. F. Llorca, and M. A. Sotelo, "The experience of DRIVERTIVE-DRIVERless cooperaTIve VEhicle-team in the 2016 GCDC," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 4, pp. 1322–1334, Apr. 2018.

[34] J. Li, C. Lu, Y. Xu, Z. Zhang, J. Gong, and H. Di, "Manifold learning for lane-changing behavior recognition in urban traffic," in *Proc. IEEE 22th Int. Conf. Intell. Transp. Syst.*, 2019, pp. 3663–3668.

[35] R. Izquierdo, A. Quintanar, I. Parra, D. Fernández-Llorca, and M. A. Sotelo, "Experimental validation of lane-change intention prediction methodologies based on CNN and LSTM," in *Proc. IEEE Intell. Transp. Syst. Conf.*, 2019, pp. 3657–3662.

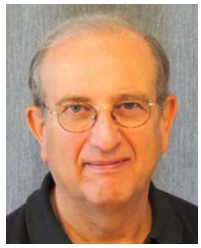
[36] R. Izquierdo *et al.*, "Vehicle lane change prediction on highways using efficient environment representation and deep learning," *IEEE Access*, vol. 9, pp. 119 454–119 465, 2021.

[37] S. Lefevre, D. Vasquez, and C. Laugier, "A survey on motion prediction and risk assessment for intelligent vehicles," *ROBOMECH J.*, vol. 1, no. 1, pp. 1–14, 2014.
[38] K. Messaoud, I. Yahiaoui, A. Verroust-Blondet, and F. Nashashibi, "Attention based vehicle trajectory prediction," *IEEE Trans. Intell. Veh.*, vol. 6, no. 1, pp. 175–185, Mar. 2021.
[39] J. Colyar and J. Halkias, "NGSIM - US highway 101 dataset," 2007. [Online]. Available: https://www.fhwa.dot.gov/publications/research/operations/07030/
[40] J. Halkias and J. Colyar, "NGSIM - Interstate 80 freeway dataset," 2006. [Online]. Available: https://www.fhwa.dot.gov/publications/research/operations/06137/
[41] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, "The highD dataset: A drone dataset of naturalistic vehicle trajectories on German highways for validation of highly automated driving systems," in *Proc. IEEE Intell. Transp. Syst. Conf.*, 2018, pp. 2118–2125.
[42] J. Bock, R. Krajewski, T. Moers, L. Vater, S. Runde, and L. Eckstein, "The IND dataset: A drone dataset of naturalistic vehicle trajectories at German intersections," 2019, *arXiv:1911.07602*.
[43] W. Zhan *et al.*, "INTERACTION dataset: An INTERnational, adversarial and cooperative moTION dataset in interactive driving scenarios with semantic maps," 2019, *arXiv:1910.03088*.
[44] H. Zhao, C. Wang, Y. Lin, F. Guillemard, S. Geronimi, and F. Aioun, "On-road vehicle trajectory collection and scene-based lane change analysis: Part I," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 1, pp. 192–205, Jan. 2017.
[45] P. Wang, X. Huang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The apolloscape open dataset for autonomous driving and its application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2702–2719, Oct. 2020.
[46] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proc. Scand. Conf. Image Anal.*, vol. 2749, 2003, pp. 363–370.
[47] S. K. He, X. Zhang, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[48] A. Quintanar, R. Izquierdo, I. Parra, D. F. Llorca, and M. A. Sotelo, "The PREVENTION challenge: How good are humans predicting lane changes," in *Proc. IEEE Intell. Veh. Symp.*, 2020, pp. 45–50.
[49] D. Wu, J. Chen, N. Sharma, S. Pan, G. Long, and M. Blumenstein, "Adversarial action data augmentation for similar gesture action recognition," in *Proc. Int. Joint Conf. Neural Netw.*, 2019, pp. 1–8.

Mahdi Biparva received the Ph.D. degree in computer science from York University, Toronto, ON, Canada, in 2019. He is currently a Research and Development Engineering with Sunnybrook Research Institute. Between 2013–2019, he was a Teaching Assistant with York University, and Research Assistant with Lab for Active and Attentive Vision, York University. His research interests include computer vision and deep learning approaches for different tasks in healthcare.

David Fernández-Llorca (Senior Member, IEEE) received the Ph.D. degree in telecommunication engineering from the University of Alcalá (UAH), Alcala de Henares, Spain, in 2008. Since November 2020, he has been a Scientific Officer with European Commission - Joint Research Center, collaborating in the HUMAINT project. He is a Full Professor (special leave) with UAH and the Co-Head of the Intelligent Vehicles and Traffic Technologies research group. He has authored more than 130 publications and more than 10 patents. His research interests include trustworthy AI for autonomous vehicles, predictive perception for autonomous vehicles, human-vehicle interaction, end-user oriented autonomous vehicles, and assistive intelligent transportation systems. He was the recipient of the IEEE ITSS Young Research Award in 2018 and IEEE ITSS Outstanding Application Award in 2013. He is currently the Editor-in-Chief of the *IET Intelligent Transport Systems*. He was an Associate Editor for the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS during 2012–2020, and *Journal of Advanced Transportation* during 2016–2020. He was the Program Chair of the IEEE ITSC 2019.

Rubén Izquierdo Gonzalo received the M.S. and Ph.D. degrees in industrial engineering from the University of Alcalá (UAH), Alcala de Henares, Spain, in 2018 and 2020, respectively. He is currently a Postdoc Researcher with Intelligent Vehicles and Traffic Technologies group. His research interests include predictive prediction for autonomous vehicles and cooperative autonomous driving. He was the main Developer of the DRIVERTIVE team that was the recipient of the Best Team with Full Automation in the Grand Cooperative Driving Challenge 2016. He was the recipient of the Social Transfer Council Award UAH in 2018.



John K. Tsotsos (Fellow, IEEE) received the Ph.D. degree in computer science from the University of Toronto, Toronto, ON, Canada. He is currently a Distinguished Research Professor of vision science with York University, Toronto, ON, Canada. He was on the Faculty with the Department of Computer Science in 1980, where he founded the university's Computer Vision Group, which he led for 20 years. In 2000, he was recruited to York University as the Director of the Centre for Vision Research. His research interests include comprehensive theory of visual attention in humans. A practical outlet for this theory embodies elements of the theory into the vision systems of mobile robots. He is a Fellow of the Royal Society of Canada, has been a CIfAR Fellow, and was the recipient of several paper prizes and other awards.